

Information Processing Society of Japan
33rd (1986, Second Term) National Assembly

A SUPPORT EXPERT SYSTEM FOR SECONDARY STRUCTURAL
PREDICTIONS IN PROTEINS

5

Kayori FURUTA, Masaaki KATOH, Tutomu SHIINO
OKI Elec. Ind. Co., Ltd. OKI Techno Systems Laboratory, Inc.

10 1. Introduction

It is known that a property of a protein largely depends on a three-dimensional structure (a tertiary structure). Accordingly, it is necessary to synthesize proteins having desired tertiary structures in order to artificially obtain a protein having a desired property. At present, however, no method is proposed for accurately grasping what tertiary structure a protein having an arbitrary amino acid sequence (a primary structure) has. For this reason, based on data obtained by a protein structure analysis (an analysis based on an X-ray diffracted image, an analysis based on magnetic resonance, or an analysis based on an energy calculation method) carried out to limited proteins which are relatively low in molecular weight, a correlation between an amino acid sequence and the structure is grasped as a statistic quantity, and the structure of the protein having a new amino acid sequence is predicted. During this method, it is proposed to calculate not a direct correlation between the primary structure and the tertiary structure but the correlation therebetween with a local structure (a secondary structure) in the amino acid sequence as an intermediate structure of the protein interposing therebetween. Several methods for predicting the

15

20

25

correlation between the primary structure and the secondary structure have been contrived. However, the number of data obtained by the three-dimensional structure analyses to date is small, and both scales and fields of proteins serving as samples are not uniform. Therefore, there is no denying that the data is insufficient to be used as statistic quantities. As a result, these prediction methods have different rules according to the manner of interpreting the data obtained by the structure analysis [JMCF85] and [SCHU79]. In the present state, an experimental verification is carried out based on prediction results obtained from these rules or a more accurate prediction method has been studied.

10 In this report, we will explain configuration and functions of a support expert system that can predict the secondary structure of a protein at a high rate using several secondary structural prediction methods which are proposed to date, and can cope with development of a new prediction method.

15 2. Protein Structural Prediction

The prediction methods explained in Chapter 1 are not complete but quite useful in that the methods can be used as indexes of the study [JMCF85].

Among the prediction methods for predicting the secondary structure from the primary structure which are contrived to date, it is considered that three methods, i.e.,
20 Chou & Fasman method, Robson method, and Nagano method are effective. Each of these methods quantitatively evaluates an expression frequency of an amino acid based on the fact that it is empirically known that the amino acid frequently appears on a specific secondary structure from the protein structure analysis result, shows an evaluation result by a value called "a trend index", and predicts the secondary structure
25 based on the trend index [SCHU79] and [Nagano85].

The following two problems, which should be considered, arise. (1) Uncertainties of prediction methods themselves cannot be determined only by individually applying the respective prediction methods. (2) There is a probability that not only the trend index changes according to an increase in three-dimensional structure analysis result but also a prediction scheme itself changes by discovery of a new rule. These are notable, characteristic problems with research support systems of this type.

3. Protein Secondary Structural Prediction Support Expert System

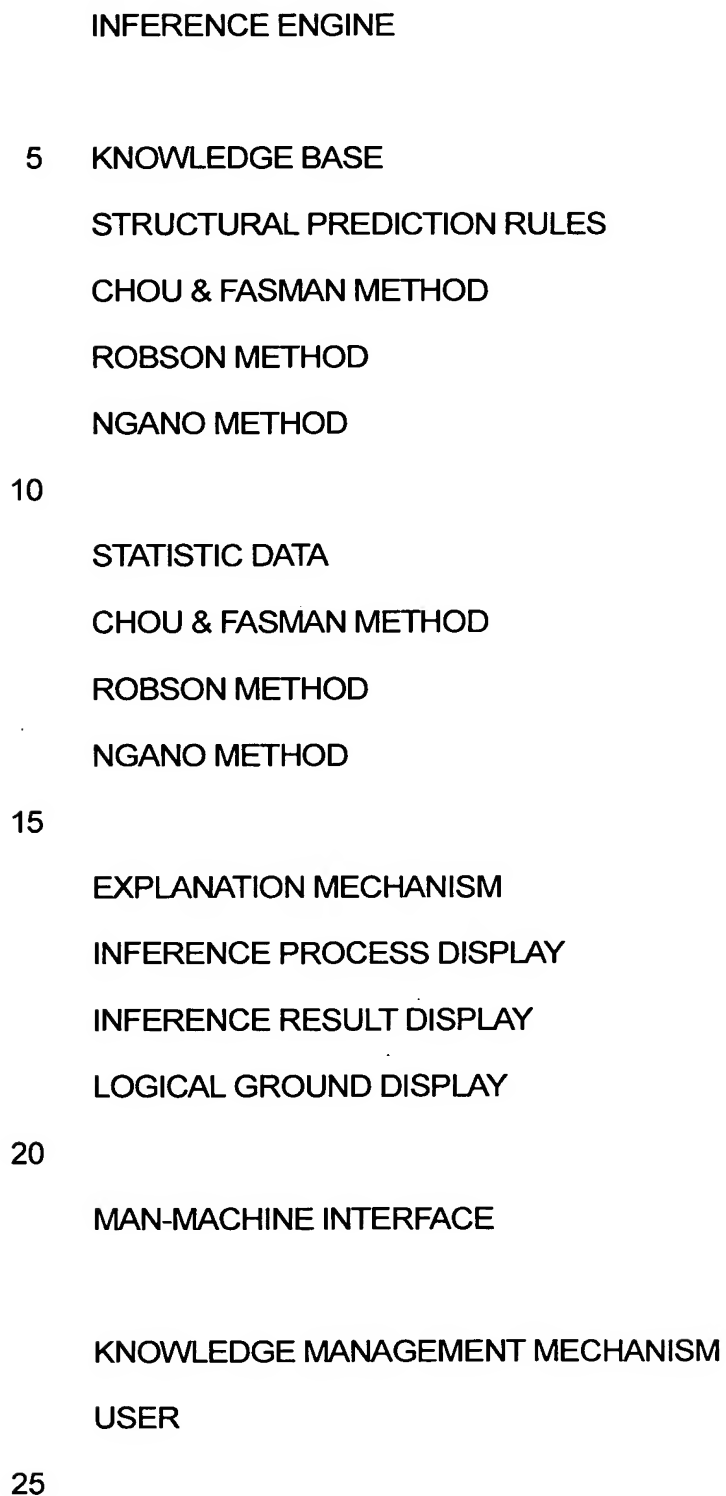
In order to handle these problems, the following functions are required. (1) A function of simultaneously applying a plurality of prediction methods, and relatively comparing the uncertainty of each prediction method with that of the other prediction method. (2) A function of coping with fluidity of knowledge, i.e., addition, change, or improvement in precision of the knowledge.

We have, therefore, constructed a prototype of a protein secondary structural prediction support expert system having these functions on an AI workstation if1000. The present system collects a plurality of prediction methods, and realizes, as independent modules, knowledge on a knowledge base [Yamamoto86]. The present system presents results including the uncertainties of the respective prediction methods, elucidates coincident parts and contradictory parts of the inference results, explains a logical ground for each inference result, and each inference process, and thereby supports thinking of each researcher.

3.1 System Configuration and Functions

Fig. 1 depicts logical configuration of the system.

Fig. 1 System logical configuration



The present system is composed of a knowledge base which stores necessary knowledge for the secondary structural prediction, an inference engine that controls an inference using this knowledge base, an explanation mechanism that outputs inference processes, inference results, and logical grounds for the respective inference results, a knowledge management mechanism that controls addition and correction of the knowledge, and a man-machine interface unit.

The knowledge base is composed of two independent modules: (1) a structural prediction rule formulating unit that formulates a plurality of prediction methods into rules, and (2) a statistic data unit that describes statistically probabilistic data calculated by the structure analyses. At present, the structural prediction rule formulating unit (1) formulates the three prediction methods (Chou & Fasman method, Robson method, and Nagano method) explained in Chapter 2 into rules.

The explanation mechanism has (1) explanation functions of explaining the inference processes and the logical grounds for the respective inference results, and (2) a display function of displaying each inference result as an image.

The knowledge management mechanism has a function of carrying out correction of the knowledge already stored in the knowledge base, addition of a knowledge to the knowledge base, and the like.

The man-machine interface unit controls a keyboard and a mouse as user input, and controls a display as user output.

As the inference engine, if1000 Prolog/SI interpreter is used.

3.2 System Output

Figs. 2 and 3 depict examples of outputs of the present system.

Fig. 2 depicts the inference result obtained by the Chou & Fasman method.

In Fig. 2, a secondary structure region (alpha-helix, beta-sheet, and bent structure) as the inference result is typically expressed, relative to an input target amino acid sequence.

Fig. 3 depicts a comparison of the inference results based on the three prediction methods explained in Chapter 2. In Fig. 3, coincident parts and contradictory parts of the secondary structure region obtained from the three inference results for the input target amino acid sequence are elucidated.

Fig. 2 Display of inference prediction result by Chou & Fasman method (α helix, β sheet, bent structure)

AMINO ACID SEQUENCE

SECONDARY STRUCTURE

TREND INDEX

Fig. 3 Display of comparison of inference results (C: Chou & Fasman method, R: Robson method, N: Nagano method)

4. Embodiment of Explanation Mechanism

An embodiment of the expert system in respect of the explanation function required by the system will be explained.

4.1 Explanation of Inference Process

The structural prediction rules in the knowledge base used for the inference are displayed in order of ignition. The user can verify correctness of the inference by viewing the displayed rules.

4.2 Explanation of Logical Ground for Inference Result

Hypotheses (including an initial hypothesis) which are built up during the inference process are sequentially displayed until the inference result is obtained.

- 5 Details of the indexes (statistically probabilistic quantities including the uncertainty) that change in the inference process are displayed. From this display result, a chain of what hypotheses are generated in the inference process based on each prediction method, and which hypotheses disappear based on indexes given to the respective hypotheses is logically clarified. As a result, each ground for which the final inference
- 10 result is led is made clear, and a difference in logic among the respective prediction methods become clear. The user determines what prediction is most probable while viewing these logical grounds.

5. Conclusion

- 15 The protein secondary structural prediction support expert system includes the knowledge base that collects a plurality of prediction methods, and has the function of explaining the logical ground for each prediction method. The present system provides a criterion for predicting structures of proteins the structures of which are unclear and those of unknown proteins, and contributes to initial value setting for
- 20 detailed structure analysis. In the future, we will intensify the functions of the knowledge management mechanism, the explanation mechanism, and the like in the prototype currently under development, and particularly expand the man-machine interface such as a knowledge base editor.

- Further, as for direction of future system development, if various pieces of data
- 25 are sufficiently available following the development of protein engineering, then we

expect that the present system can be used as a support system (having the logical explanation function, etc.) for contriving a new prediction method, and that the present system evolves into a more accurate structural prediction expert system by implementing the contrived new prediction method into the present system.

5

ACKNOWLEDGEMENT

Sincere thanks to Professor Kenichi TOMITA of Osaka University, and Professor Noritake YASUOKA of Himeji Institute of Technology for supervision of our study.

10

Reference Literature

[SCHU79] SCHULZ, et al., "Principles of Protein Structure" Springer-Verlag, 1979.

[JMCF85] The Japan Machinery Federation and others, "Future Industrial Infrastructure Technology Research Report, 1985"

15

[Nagano85] NAGANO "Protein Engineering", pp. 30 to 52, CMC Publishing Co., Ltd., 1985

[Yamamoto86] Yamamoto, et al., "Uncertainty Inference of Secondary Structural Prediction Method", Information Processing Society of Japan, 33rd National Assembly Journals.

20



第33回(昭和61年後期)全国大会

講演論文集 (II)

データベース

ネットワークおよび分散処理

パターン処理および人工知能

自然言語処理

昭和61年10月 1 日～ 3 日



社団法人

情報処理学会

Information Processing Society of Japan

蛋白質の二次構造予測支援エキスパート・システム

2L-8

古田 香代里 * 加藤 正明 ** 椎野 努 *
* 沖電気工業(株) ** 沖テクノシステムズラボラトリ(株)

1. はじめに

蛋白質の性質は、その立体構造(三次構造)に強く依存することが知られている。従って、希望する性質の蛋白質を人工的に得るためには、希望する三次構造の蛋白質を合成する必要がある。しかし現在、任意のアミノ酸配列(一次構造)を持つ蛋白質がどのような三次構造を持つかを正確に知る方法は存在しない。そこで、限定された、しかも比較的分子数の少ない蛋白質の立体構造解析(X線回折像による解析や磁気共鳴による解析、エネルギー計算法による解析)によって得られたデータを基に、アミノ酸配列と構造との相関関係を統計的な量として捉え、それによって新しいアミノ酸配列の蛋白質の構造を予測する方法がとられる。この際、一次構造と三次構造の間の直接的な相関を求めるのではなく、その中間構造であるアミノ酸配列中の局部的構造(二次構造)を仲介にした相関を求めることが提唱され、一次構造と二次構造の相関を予測する方法がいくつか考案されている。しかしながら、立体構造解析によって現在までに得られているデータは、数も少なく、その試料となる蛋白質の規模、分野共に偏りがあり、統計量として扱うには不十分であることは否めない。従って、これらの予測法は、構造解析によって得られているデータの解釈のしかたによって、それぞれ異なった法則となっている [JMCF85] [SCHU79]。現状では、これらの法則から得られた予測結果を基に実験的検証が行われ、より精度の高い予測法の研究が行われている。本稿では、このような現状から、現在考案されているいくつかの二次構造予測法により、蛋白質の二次構造予測を高速に行うと共に、新しい予測方式の発展にも対処できる支援エキスパート・システムの構成と機能について述べる。

2. 蛋白質の構造予測

1章で述べた予測法は、不完全ではあるが、研究の指標となるという面では非常に有用である [JMCF85]。

現在までに考案されている一次構造から二次構造の予測法の中で、Chou & Fasman法、Robson法、Nagano法の三つの方法が有力とされている。これらの方法は、蛋白質の立体構造解析結果から、アミノ酸が特定の二次構造に多く出現する事実が経験的に知られていることを用いて、この出現頻度を定量的に評価

した結果を傾向性指数と呼ばれる値で示し、それを基準として二次構造を予測する [SCHU79] [長野85]。

ここで問題となることは以下の2点である。①個々の予測法を単独に適用するだけでは、その予測法自体の持つ不確かさを判断できない。②立体構造解析結果の増加によって傾向性指数が変化するだけでなく、新しい法則性の発見により、予測手法そのものが変化する可能性がある。これらは、この種の研究支援システムの構築にとっての大きな特徴である。

3. 蛋白質の二次構造予測支援エキスパート・システム

上述の問題に対処するためには、①複数の予測法を同時に適用し、個々の予測法の持つ不確かさを他の予測法と相対的に比較できる。②知識の流動性、すなわち、知識の追加、変更あるいは精密化に対応し得る、という機能が必要である。

そこで、これらの機能を有する蛋白質の二次構造予測支援エキスパート・システムのプロトタイプをAIワークステーション if1000上に構築した。本システムでは、複数の予測法を集結し、さらに知識を独立したモジュールとして知識ベースに実現した [山本86]。本システムは、各々の予測法による不確かさを含む結果を提示し、推論結果が一致する部分と相反する部分を明示し、その論理的根拠および推論過程を説明することにより、研究者の思考を支援する。

3.1 システムの構成と機能

システムの論理構成を図1に示す。

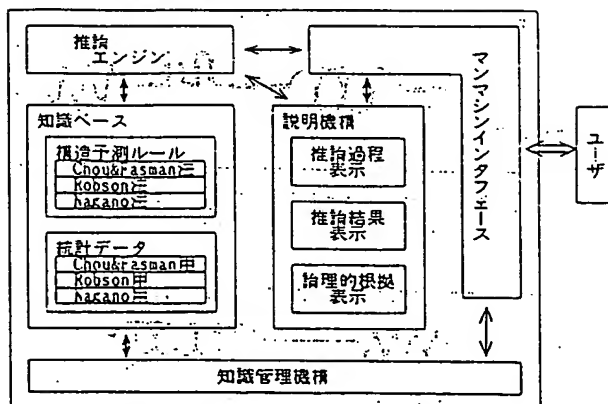


図1. システムの論理構成

An expert system for secondary structural predictions in proteins

Kayori FURUTA * Masaaki KATOH ** Totonu SHIINO *

* OKI Elec. Ind. Co., Ltd. ** OKI Techno Systems Laboratory, Inc.

本システムは、二次構造予測に必要な知識を格納した知識ベースと、この知識ベースを用いた推論を制御する推論エンジン、推論過程および推論結果とその論理的根拠を出力する説明機構、知識の追加修正を制御する知識管理機構、マンマシンインタフェース部から成る。

知識ベースは、(1) 複数の予測法をルール化した構造予測ルール部と、(2) 構造解析により算出された統計確率的データを記述した統計データ部の、2つの独立したモジュールから構成されている。現在、(1) の構造予測ルール部には2章で述べた3種類の予測法(Chou & Fasman法、Robson法、Nagano法)をルール化している。

説明機構は、(1) 推論過程説明機能と推論結果の論理的根拠説明機能、(2) 推論結果の画像表示機能を有する。

知識管理機構は、知識ベースに対して、既に格納されている知識の修正や新たな知識の追加等の機能を有する。

マンマシンインタフェース部は、ユーザの入力としてキーボード、マウスを、出力としてディスプレイを制御する。

推論エンジンはif1000のProlog/SLインタプリタを用いる。

3.2 システムの出力

本システムの出力例を図2、図3に示す。

図2は、Chou & Fasman法による推論結果を表示したものである。入力された対象アミノ酸配列に対して、推論結果の二次構造(α ヘリックス、 β シート、折れ曲り構造)の部位を模式的に表現している。

図3は、2章で述べた3種類の予測法による推論結果の対照表示である。入力された対象アミノ酸配列に対する3種類の推論結果から得られた二次構造の部位の一致する部分と相反する部分を明示している。

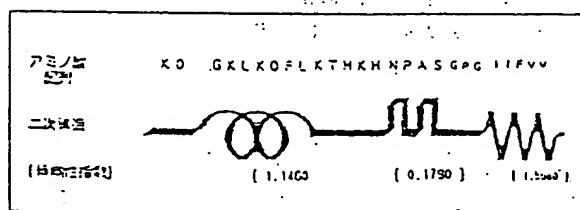


図2. Chou&Fasman法による推論予測結果表示
(α : α ヘリックス β : β シート γ :折れ曲り構造)

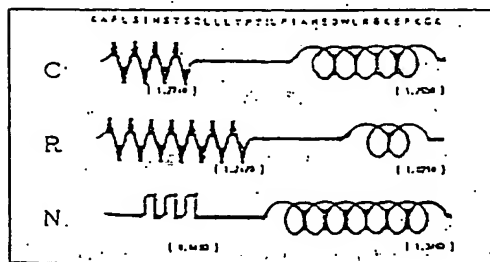


図3. 推論結果の対照表示
(C:Chou&Fasman法, R:Robson法, N:Nagano法)

4. 説明機構の実現形態

エキスパート・システムに要求される説明機能について本システムにおける実現形態を示す。

4.1 推論過程の説明

推論に用いた知識ベース中の構造予測ルールを、発火の順に表示する。ユーザは表示されたルール群を見ることで、推論の正当性を検証できる。

4.2 推論結果の論理的根拠の説明

推論過程に生じる仮説(初期状態を含む)を推論結果に至るまで順次に表示する。この時に変化する指数(不確実性を含む統計確率的な量)の経緯を示す。この表示結果から、各予測法による推論の過程でどのような仮説が生成され、仮説に与えられている指数をもとに、どの仮説が消滅したかの連鎖が論理的に明らかにされる。これによって、最終的に推論結果の導かれた根拠が明確になり、各予測法の論理の違いも明確になる。ユーザ以上の論理的根拠を見ながら、どちらがより確からしいかの判断を行うことになる。

5. おわりに

蛋白質の二次構造予測支援エキスパート・システムは、複数の予測法を集結した知識ベースを有し、各予測法の論理的根拠を説明する機能を有する。本システムは、構造の明らかでない蛋白質あるいは未知の蛋白質の構造予測に目安を与え、詳しい構造解析のための初期値設定にも寄与する。今後、現在開発中のプロトタイプ知識管理機構や説明機構等の機能強化、特に知識ベースエディタ等のマンマシンインタフェースの拡充を図る予定である。

また、今後のシステムの発展の方向としては、蛋白質工学の発展に伴い、各種データが充実してきた場合、新しい予測法の発案のための支援システム(論理的説明機能等)としても活用でき、それによって発案された新しい予測法は、そのまま本システムにインプリメントすることにより、より正確な構造予測エキスパート・システムとして進化することが期待できる。

謝辞

本研究を行うにあたり御指導を頂いた大阪大学富田研一教授、姫路工業大学安岡則武教授に感謝致します。

参考文献

- [SCHU79] SCHULZ, et al. "Principles of Protein Structure" Springer-Verlag, 1979.
- [JMCF85] (社)日本機械工業連合会 他「昭和59年度 次世代産業基盤技術調査研究報告書」, 1985.
- [長野85] 長野「プロテインエンジニアリング」pp.30-52 シーエムシー, 1985.
- [山本86] 山本 他「二次構造予測法における不確実性推論」第33回情報処理学会全国大会論文集

Volume 208

Number 1

25 April 1997

BEST AVAILABLE COPY

0073

③

D/24

JMB

JOURNAL OF MOLECULAR BIOLOGY

DO124BAR

1997:268(1)

科学技術振興事業区
科学技術情報事業本部



0022-2836(19970425)268:1;1-I

コ 無 BD I C4

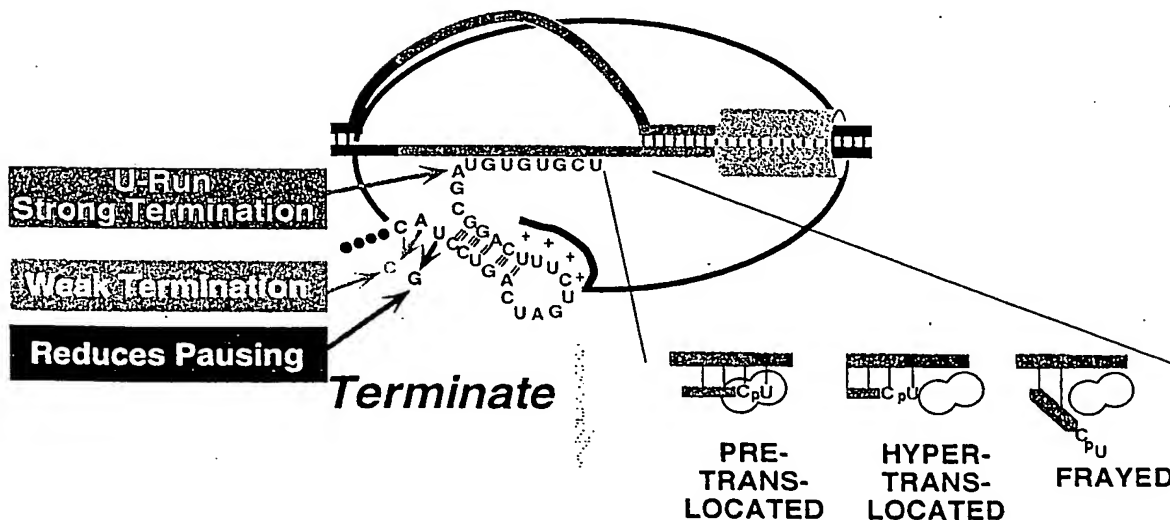
<9>970456713

1997.05.16

書誌件数	和文許諾件数	英文許諾件数



<11>19970456714



268(1)

ACADEMIC PRESS

268 (1) 1-227 ISSN 0022-2836



0022-2836(199704)268:1;1-T

IDEAL